因果関係を考慮した連鎖パタンマイニング手法のパラメータの設定

因果関係を考慮した連鎖パタンマイニング手法のパラメータの設定

SS4-7

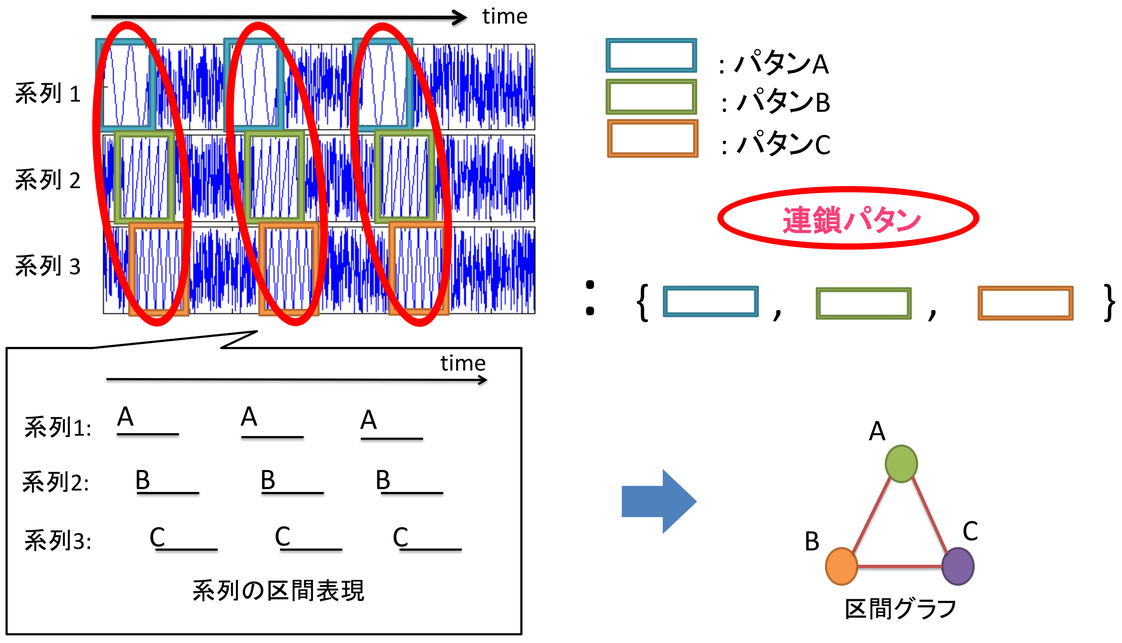
因果関係を考慮した連鎖パタンマイニング手法のパラメータ設定

○ 大久保勇輔　李セロン　岡田吉史　（室蘭工業大学）

背　景

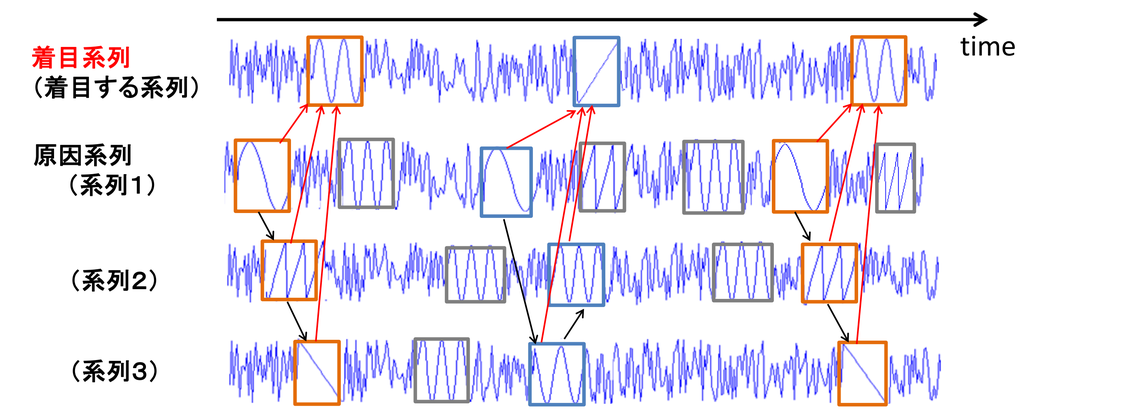
▽連鎖パタンマイニング手法

複数の系列間にまたがって，同時刻帯に繰り返し現れる頻出パタン(連鎖パタン)を抽出する手法　(*2012. miura*)



本研究の目的

以前の連鎖パタンマイニング手法を拡張して、
複数の系列データ（脳波、脈波など）から、興味の事象の発生原因を見つけ出す



解析者が着目する系列（着目系列）に、出現するパタン（結果パタン）があると仮定

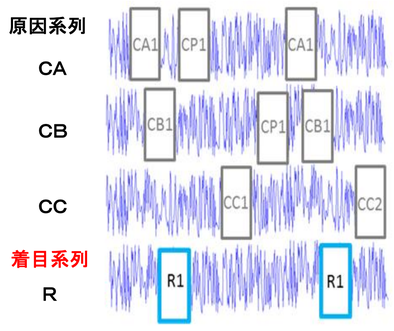
結果パタンの誘導原因となるパタン（原因パタン）をその他の系列から特定

本研究の応用例
- 複数のバイタルデータ（脳波、心電図）から、病理的特徴を抽出
- バイタルデータ、音声データから感性情報を解析

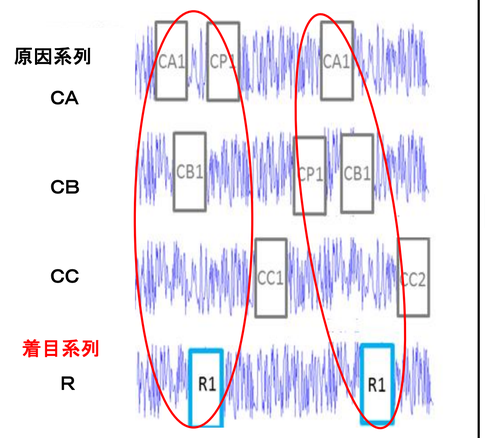
手　法

□手法の流れ

Input: 系列データ



Step1：連鎖パタンを抽出

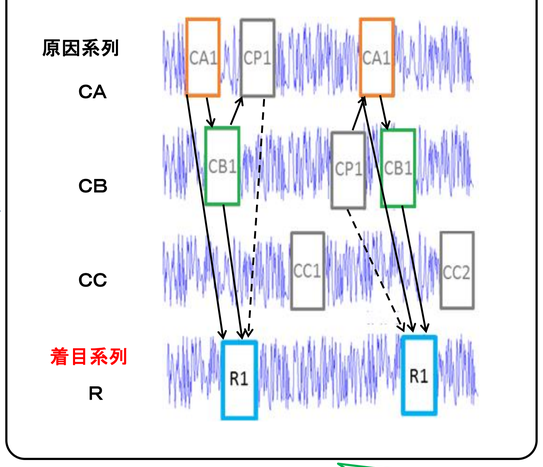


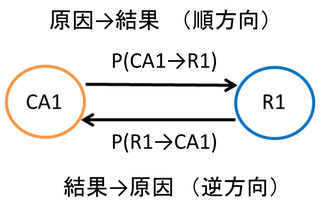
・前手法を用いて連鎖パタンマイニングを行う

- 同時刻帯に重なって出現する頻出パタンを連鎖パタンとして抽出

連鎖パタン：　{CA1,CB1,R1,CP1}
　　{CP1,CA1,CB1,R1}

Step2：因果の確率を計算





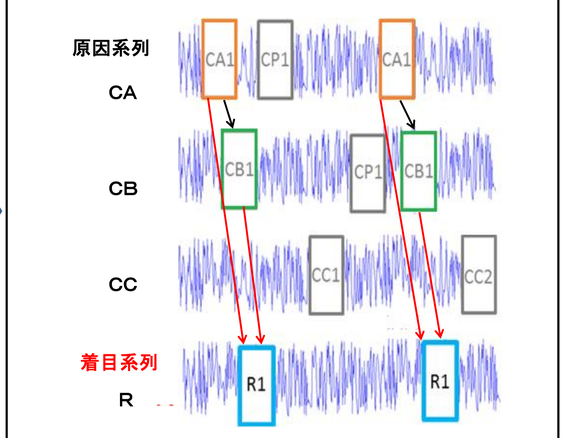
・CA1が発生した条件下でR1が発生する確率

T1：　順方向の確率の閾値

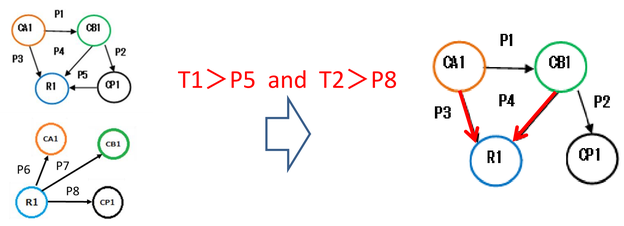
・R1が発生した条件下でCA1が発生する確率

T2：　逆方向の確率の閾値

Step3：因果関係のないエッジを削除



原因パタンと結果パタンのエッジに付与された確率がT1より下回るかつ、T2より下回る場合、エッジを削除する



Output:

原因パタン：CA1,CB1
結果パタン：R1

結果パタンと接続されている原因パタンを出力

評価実験

□ 人工データセットを用いて抽出精度を評価
- 一様乱数による系列データに連鎖パタン（正解パタン）を2種類とノイズとなる頻出パタンを埋め込んだ

□　評価指標

適合率($Precision$) = $\frac{CDP}{DDP}$　　再現率($Recall$) = $\frac{CDP}{EDP}$

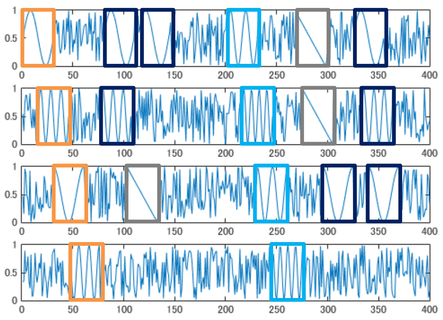
F値 = $\frac{2\times Precision\times Recall}{Precision+Recall}$

CDP　：　正しく抽出された正解パタンのデータ点数
DDP　：　本手法により連鎖パタンとして抽出された部分のデータ点数
EDP　：　正解パタンのデータ点の総数

□　データセットの種類

種類	データセットの内容
dataset1	パタンの長さを変更　(dataset1.1, dataset1.2, dataset1.3)
dataset2	正解パタンの出現間隔と各系列の長さを変更 (dataset2.1, dataset2.2, dataset2.3)
dataset3	正解パタンの出現頻度を変更　(dataset3.1, dataset3.2, dataset3.3)
dataset4	ノイズパタンの出現頻度を変更 (dataset4.1, dataset4.2, dataset4.3)
dataset5	ノイズパタンに正解パタンの一部（ダミーパタン）を挿入し、ダミーパタンの出現頻度を変更 (dataset5.1,dataset5.2,dataset5.3)

□　人工データの一部

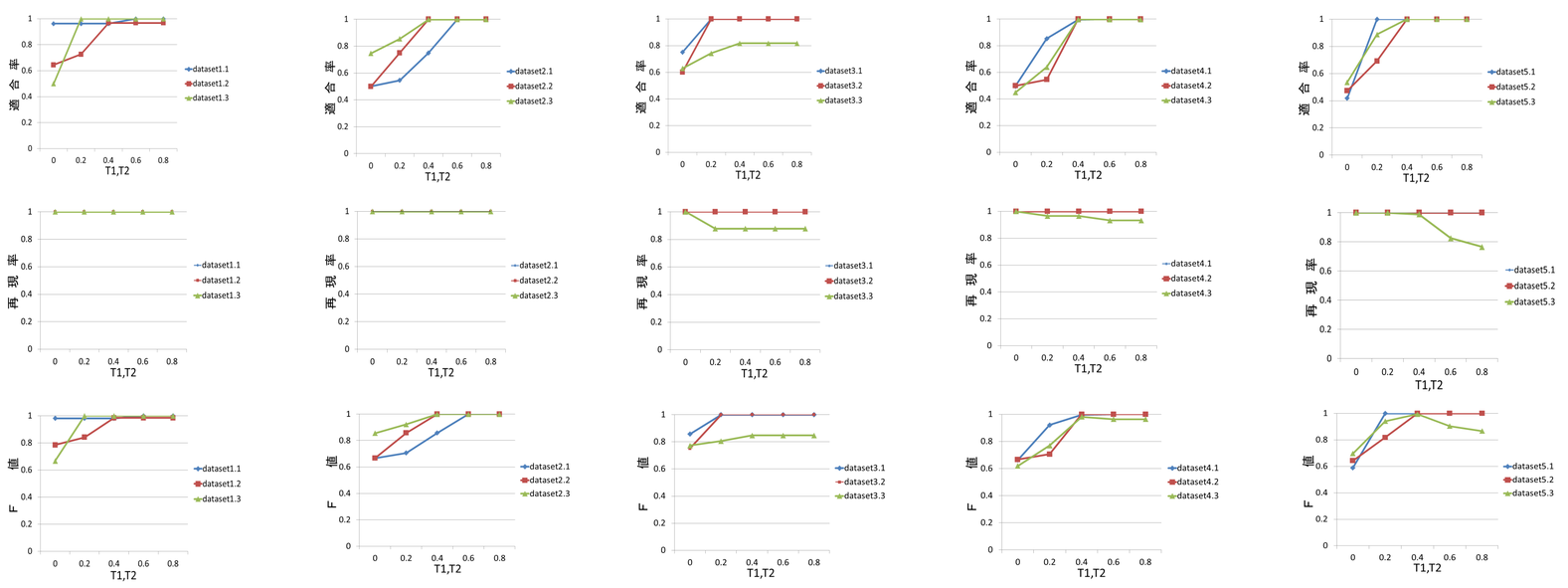


正解パタン

ノイズパタン

ダミーパタン

結果と考察



○ dataset1～3, 4.1,4.2,5.1,5.2
T1,T2の値が大きいほどF値は高くなる
⇒ 結果パタンとは無関係な原因パタンを連鎖パタンから削除することにより、因果関係のあるパタンで構成される連鎖パタンのみ正しく抽出している

○ dataset4.3, 5.3 (ノイズパタン、ダミーパタンの出現頻度が大きい場合)
T1,T2= 0.4のとき、F値がピークとなりそれより大きなT1,T2では抽出精度が低下
⇒ノイズパタン、ダミーパタンの影響を受けて正解パタンを正しく検出できなくなる

○ dataset2.1, 3.3 (正解パタンの出現間隔が短い or 出現頻度が大きい場合)
T1,T2が低く設定されたとき、他のデータセットより抽出精度が低下
⇒ パタンが密に埋め込まれ、ノイズパタンが混入しやすくなるため

まとめ

○パタンが密に埋め込まれているデータセットでは、T1,T2の値を大きく設定することにより、抽出精度が高くなる

○ダミーパタンが含まれたデータセットは、知見が得られなかった

□今後の課題
- 系列データの性質に応じて、閾値を設定する方法の開発
- 実データへの適用実験を行い、よりノイズに頑健な手法へと拡張を目指す